Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study

Fumiyo Nakatsuhara, Chihiro Inoue, *CRELLA, University of Bedfordshire*

Vivien Berry, *British Council*

Evelina D. Galaczi, *Cambridge English Language Assessment*

ABSTRACT

This research explores how internet-based video-conferencing technology can be used to deliver and conduct a speaking test, and what similarities and differences can be discerned between the standard and computer-mediated face-to-face modes. The context of the study is a high-stakes speaking test, and the motivation for the research is the need for test providers to keep under constant review the extent to which their tests are accessible and fair to a wide constituency of test takers. The study examines test-takers' scores and linguistic output, and examiners' test administration and rating behaviors across the two modes. A convergent parallel mixed-methods research design was used, analyzing test-takers' scores and language functions elicited, examiners' written comments, feedback questionnaires and verbal reports, as well as observation notes taken by researchers. While the two delivery modes generated similar test score outcomes, some differences were observed in test-takers' functional output and the behavior of examiners who served as both raters and interlocutors.

KEYWORDS

Speaking assessment, Video-conferencing technology, Mixed-methods research

ACKNOWLEDGEMENTS & FUNDING

We acknowledge the participation of Lynda Taylor in the design of both Examiner and Test-taker Questionnaires, and Jamie Dunlea for the FACETS analysis of the score data. Our special thanks go to Jermaine Prince for his technical support, careful observations and professional feedback. This research was funded by the IELTS Partners: British Council, Cambridge English Language Assessment and IDP (IELTS) Australia.

INTRODUCTION

Face-to-face interaction no longer depends upon physical proximity within the same location, as recent technical advances in online video-conferencing technology have made it possible for users in two or more locations to successfully communicate in real time through audio and video. Video-conferencing applications such as Skype and Facetime are now commonly used to communicate in personal or professional settings when those involved are in different locations. The use of video-conferencing is also prevalent in educational contexts, including second/foreign language (L2) learning (e.g., Abrams, 2003; Smith, 2003; Yanguas, 2010). Video-conferencing in L2 speaking assessment is less widely used, and research on this test mode is scarce, with notable exceptions being studies by Clark and Hooshmand (1992), Craig

and Kim (2010), and Kim and Craig (2012). The present study aims to extend the research base on the use of video-conferencing in L2 speaking assessment through an exploration of test-takers' scores and linguistic output, and examiners' test management and rating behaviors, across two different delivery modes: a standard face-to-face mode and a video-conferencing mode. The context of use is the speaking component of the International English Language Test System (IELTS) test, which is a high-stakes, four-skills test of English language proficiency typically used for educational, professional and migration purposes. The on-line video-conferencing program, Zoom (www.zoom.us), was used to administer the video-conferencing version of the test, as it is considered to be a more stable computer-mediated communication software than other programs such as Skype (see Nakatsuhara, Inoue, Berry & Galaczi, 2016 for the detailed rationale for selecting this software).

The research was motivated by the need for test providers to keep under constant review the extent to which their tests are accessible and fair to as wide a constituency of test users as possible. Face-to-face tests for assessing spoken language ability offer many benefits, particularly the opportunity for reciprocal interaction. However, face-to-face speaking test administration is usually logistically complex and resource-intensive, and the face-to-face mode may therefore be impossible to conduct in geographically remote or politically unstable areas. An alternative in such circumstances could be to use a semi-direct speaking test where the test-taker speaks in response to recorded input, usually delivered by computer. A disadvantage of this approach is that the delivery mode precludes reciprocal interaction between speakers, thus constraining the test construct.

It is appropriate, therefore, to explore how new technologies can be harnessed to deliver and conduct the face-to-face version of an existing speaking test, and to discern what similarities and differences between the two modes exist. Such an exploration holds the potential for a practical, theoretical and methodological contribution to the L2 assessment field. First, it contributes to an under-researched area which, due to technological advances, is now becoming a viable possibility in speaking assessment and therefore provides an opportunity to collect validity evidence supporting the use (or not) of the video-conferencing mode as a parallel alternative to the standard face-to-face variant. Second, such an investigation could contribute to theoretical construct-focused discussions about speaking assessment in general. Finally, the investigation presents a methodological contribution through the use of a mixed-methods approach which integrates quantitative and qualitative data.

RESEARCH BACKGROUND

Role of test mode in speaking assessment

Face-to-face speaking tests have been used in L2 assessment for over a century (Weir, Vidakovic, & Galaczi, 2013) and in the process have been shown to offer many beneficial validity considerations, such as an underlying interactional construct and positive impact on learning. However, they are constrained by low practicality due to the 'right-here-right-now' nature of face-to-face tests and the need for the development and maintenance of a worldwide cadre of trained examiners. The resource-intensive demands of face-to-face speaking tests have given rise to several more practical alternatives, namely semi-direct speaking tests (involving the elicitation of test-taker speech with machine-delivered prompts and scoring by human raters) and automated speaking tests (both delivered and scored by computer). With

several different test modes aiming to tap into communicative speaking ability, a fundamental question to ask is whether, and/or how, the delivery medium changes the nature of the construct being measured. Despite research which has reported overall score and difficulty equivalence between computer-delivered and face-to-face tests and, by extension, construct comparability (Bernstein, Van Moere, & Cheng, 2010; Kiddle & Kormos, 2011; Stansfield & Kenyon, 1992), theoretical discussions and empirical studies which go beyond sole score comparability have highlighted the fundamental construct-related differences between different test formats. Essentially, semi-direct and automated speaking tests are underpinned by a *psycholinguistic* construct, which places emphasis on the cognitive dimension of speaking, as opposed to the *socio-cognitive* construct of face-to-face tests, where speaking is seen both as a cognitive trait and a social, interactional one (Galaczi, 2010; McNamara & Roever, 2006; van Moere, 2012). Studies (Hoejke & Linnell, 1994; Luoma, 1997; O'Loughlin, 2001; O'Sullivan, Weir & Saville, 2002; Shohamy, 1994) have also highlighted differences in the language elicited in different formats.

Differences between different speaking test formats have also been reported from a cognitive validity perspective, since the choice of format impacts the cognitive processes which a test can activate. Field (2011) notes that interactional face-to-face formats entail processing input from interlocutor(s), keeping track of different points of view and topics, and forming judgements in real-time about the extent of accommodation to the interlocutor's language. These kinds of cognitive decisions impose different processing demands on test-takers in face-to-face and computer-delivered test environments.

Test-takers' perceptions have also been found to differ according to test format, with research (Clark, 1988; Kenyon & Malabonga, 2001; Stansfield, 1990) indicating that test-takers report a sense of nervousness and lack of control when taking a semi-direct test – what we might call the 'bulldozer effect' in that the test-taker's role is controlled by the machine which cannot offer any support in cases of test-taker difficulty. It is also notable that if a group of test-takers expresses a significantly stronger preference for one mode over another, they seem to be in favor of the face-to-face mode (Kiddle & Kormos, 2011; Qian, 2009).

Video conferencing and speaking assessment

The choice of speaking test format is therefore not without theoretical and practical consequences, as the different formats offer their own unique advantages, but inevitably come with certain limitations. As Qian (2009, p.124) reminds us in the context of a computer-based speaking test:

> This technological development has come at a cost of real-life human interaction, which is of paramount importance for accurately tapping oral language proficiency in the real world. At present, it will be difficult to identify a perfect solution to the problem but it can certainly be a target for future research and development in language testing.

Such a development in language testing can be seen in recent technological advances which involve the use of video-conferencing in speaking assessment. This new mode preserves the co-constructed nature of face-to-face speaking tests while offering the practical advantage of

remotely connecting test-takers and examiners who could be continents apart. As such, it reduces some of the practical difficulties of face-to-face tests while preserving the interactional construct of this test format.

The use of a video-conferencing system in English language testing is not a recent development. In 1992 a team at the Defense Language Institute Foreign Language Center (USA) conducted an exploratory study of 'screen-to-screen testing', i.e., testing using video-conferencing (Clark & Hooshmand, 1992). The study was enabled by technical developments at the Foreign Languages Centre at the U.S. Defense Language Institute, such as the use of satellite-based video technology which could broadcast and receive in (essentially) real-time both audio and video. The technology had been mostly used for language instruction, and the possibility for incorporating it in assessment settings was explored in the study. The focus was a comparison of the face-to-face and video-conferencing modes in tests of Arabic and Russian. The researchers reported no significant difference in performance in terms of scores, but did find an overall preference by test-takers for the face-to-face mode; no preference for either test mode was reported by the examiners.

In two more recent studies, Craig and Kim (2010) and Kim and Craig (2012) compared the face-to-face and video-conferencing mode with 40 English language learners whose L1 was predominantly Korean. Their data comprised analytic scores on both modes (on fluency, functional competence, accuracy, coherence, interactiveness) and also test-taker feedback on 'anxiety' in the two modes, operationalized as 'nervousness' before/after the test and 'comfort' with the interviewer, test environment and speaking test (Craig & Kim, 2010:17). The results showed no statistically significant difference between global and analytic scores on the two modes, and the interview data indicated that most test-takers 'were comfortable with both test modes and interested in them' (Kim & Craig, 2012, p.268). The authors concluded that the video-conferencing mode displayed a number of test usefulness characteristics (Bachman & Palmer, 1996), including reliability, construct validity, authenticity, interactiveness, impact and practicality. In terms of test-taker anxiety, a significant difference emerged, with anxiety before the face-to-face mode found to be higher.

Validation of a video-conferencing test mode

The research reviewed so far has indicated that test mode affects a range of test qualities which impact on the underlying test construct and has implications for the validity argument of a test. The remote face-to-face format has the potential to optimize strengths and minimize shortcomings of existing formats by blending technology and face-to-face assessment. Its advantages and limitations have been investigated by a very small number of studies and are, as such, still an open empirical question. The present study aims to provide an exploration of the features of this new and promising speaking test format and will do so through the prism of test validity. For this purpose it will adopt the socio-cognitive validity framework for speaking tests initially proposed by Weir (2005) and further elaborated in Taylor (2011). This validity framework was selected since it was seen to provide a comprehensive, transparent and useful approach to investigating validity and, most importantly, identified the type of evidence needed for the different aspects of validity (O'Sullivan & Weir, 2011). Essentially, the current study focuses on the 'criterion-related validity' aspect of the framework, which taps into comparisons of different versions of the same test and into equivalence of parallel test versions. Such an investigation, Weir (2005) holds, needs to be based on quantitative and qualitative equivalence. Additionally, the investigation reported here explores context validity parameters, which relate to the task input and expected output, and scoring validity

parameters, which relate to ensuring that the outcomes/scores of the test are fair and meaningful, partly through considering examiner behavior and attitudes. The findings from previous investigations of the video-conferencing mode (Clark & Hooshmand, 1992; Craig & Kim, 2010; Kim & Craig, 2012) can also be said to have investigated the criterion-related and scoring aspect of validity, with their focus on criterial parameters which may distinguish (or not) across test modes and on test scores. However, they did not explicitly focus on context validity parameters such as functional language or examiner perceptions of the modes, which is an area this study hopes to further extend.

RESEARCH QUESTIONS

The study addressed the following four research questions.

Comparing the standard face-to-face and video-conferencing mode:

RQ1: Are there any statistically significant differences in test-takers' scores?
RQ2: Are there any differences in linguistic output, specifically types of language function, elicited from test-takers?
RQ3: Are there any perceived differences in examiners' test administration behavior?
RQ4: Are there any perceived differences in examiners' rating behavior?

METHODOLOGY

Research design

The study used a convergent, parallel mixed-methods design (Creswell & Plano Clark, 2011), where quantitative and qualitative data were collected in two parallel strands, and separately analyzed, after which findings were integrated. The two data strands provided different types of information and allowed for an in-depth and comprehensive set of findings. Figure 1 presents information on the data sources and analysis strands in the research design[1]. In the figure, each of the data sources and analysis methods is related to relevant research questions.

INSERT FIGURE 1 HERE

Participants

Thirty-two test-takers attending IELTS preparation courses at a London college participated in the study; 14 were male (43.8%) and 18 were female (56.3%), ranging in age from 19 to 51 years (median=28.0), with 21 different first languages. Their face-to-face IELTS Speaking

[1] This paper presents selected parts of the larger research reported in Nakatsuhara *et al.* (2016). In addition to the data sources listed in Figure 1, the full report includes test-takers' responses to feedback questionnaires and semi-structured interviews, and examiners' questionnaire responses to 27 closed-ended questions regarding their perceptions of the two test delivery modes. However, a discussion of these components is beyond the scope of this paper.

Bands ranged from 5.0 and 8.5 (mean=6.55, SD: 0.88)[2]. This sample is considered to be sufficiently representative of the overall IELTS population, since 19 out of the 21 participant L1s are in the typical IELTS top 50 test-taker L1s and 98% of the IELTS population receive scores between 4.5 and 8.5 (www.ielts.org). Four certificated, experienced IELTS examiners (Examiners A-D) also participated in the research.

Data sources

Speaking test performances and questionnaire completion. All 32 test-takers took both face-to-face and video-conferencing-delivered tests in a counter-balanced design. Two versions of the IELTS Speaking test (Versions 1 and 2) were selected to ensure comparability of tasks[3] (e.g., topic familiarity, expected output). In order to minimize any possible version effects, the order of the two versions was also counter-balanced.

The IELTS Speaking test has a one-on-one examiner/test-taker format. The test lasts 11–14 minutes and consists of three parts: Part 1 – Introduction and interview (4-5 minutes); Part 2 – Test-taker long turn (3-4 minutes), and Part 3 – Examiner and test-taker discussion (4-5 minutes)[4]. Test-takers are given a set of analytic scores (described below) for their entire performance throughout the three tasks, rather than separate scores for each part of the test.

Data collection was carried out over four consecutive days. Two examiners conducted four test sessions in both modes of delivery each day, giving a total of eight tests per examiner, per day. All test sessions were audio- and video-recorded using digital audio recorders and external video cameras. The video-conferencing-delivered test sessions were also video-recorded using Zoom's on-screen recording technology. Additionally, all test-takers took part in short semi-structured interviews and completed feedback questionnaires.

For the video-conferencing test delivery, an Apple iPad Air tablet with a wireless portable speaker was set up in both examiner and candidate rooms, and both computers were located on the same local network. The average wireless Internet speeds for downloading and uploading in the research venue were 48.9Mbit/s and 6.42Mbit/s, respectively, which met the minimum requirements for Zoom (i.e., 512kbit/s for both downloading and uploading). A technical advisor was on site to handle and record technical issues encountered during the data collection (for the resulting technical report, see Nakatsuhara *et al*., 2016, Appendix 8).

[2] The score distribution was slightly positively skewed and slightly more peaked than a Gaussian distribution (Shapiro-Wilk test: W=0.919, df=32, p=0.02).This influences the choice of statistical tests in SPSS, but FACETS makes no assumptions about data distributions (M. Linacre, personal communication, November 16, 2016).

[3] The comparability of the two versions was supported by Mann Whitney's U tests, which showed no significant differences in any analytic or overall scores, and by the analysis of observers' notes that did not identify any differences in examiner behavior. The use of a few language functions was found to be significantly different between the versions (e.g., Part 3 Asking for clarification in the face-to-face test (Z=-2.118, p=0.034); Part 1 agreeing in the video-conferencing test (Z=-2.932, p=0.003)). However, the counter-balancing of the two versions between the two delivery modes is considered to have minimized such version effects.

[4] More information on the IELTS Speaking test tasks is available at http://takeielts.britishcouncil.org/prepare-test/understand-test-format/speaking-test.

Observers' field notes. Three researchers were present in different test rooms and took field notes. All have PhDs specializing in assessing speaking and have extensive experience in data collection including observation. Two of the researchers observed test sessions in both face-to-face and video-conferencing examiner rooms. Each of them followed one particular examiner on each day, to enable them to observe the same examiner's behavior under both test delivery conditions. The research design ensured that the researchers could observe all four examiners on different days. The third researcher stayed in the video-conferencing test-taker room, so that all test-takers and examiners performing under the video-conference condition could be observed.

Examiners' ratings. Examiners awarded scores on a scale of 1-9 on four analytic rating categories: *fluency and coherence, lexical resource, grammatical range and accuracy, pronunciation*, (hereafter referred to as *fluency, lexis, grammar* and *pronunciation*) according to the standard assessment criteria and rating scales used in operational IELTS tests.[5]

Various options for carrying out multiple ratings using video-recorded performances were considered during the research planning stage. However, this could introduce a significant confounding variable, namely rating a video-recorded performance on face-to-face and video-conferencing delivery modes, whose effect we were not able to predict at that stage, due to the lack of research in this area. Given the preliminary and exploratory nature of the current research, we decided to limit this study to the use of live rating scores obtained following a rigorous counter-balanced data-collection matrix. Nevertheless, as will be described in the final section, our follow-up study involves a multiple rating design on the basis of the comparability established between live and video-recorded rating modes, as recently discussed in Nakatsuhara, Inoue and Taylor (2017).

Examiners' written notes. The rating sheet used in this study included a space under each of the four analytic scores in which examiners were asked to make brief notes to provide rationales for each of the scores that they awarded. Compared with the verbal report methodology (described below), a written description is likely to be less informative. However, given the ease of collecting large datasets in this manner, obtaining brief notes from examiners to supplement a small quantity of verbal report data was considered valuable (a data collection strategy also used in Isaacs, 2010).

Examiner feedback questionnaires. After finishing all tests each day, examiners completed a feedback questionnaire about their behavior as interlocutors and as raters under video-conferencing and face-to-face test conditions. The questionnaire consisted of 22 Likert scale questions, 5 multiple-choice questions and five free comment boxes to elaborate on the closed-question responses. Since a discussion of the full questionnaire is beyond the scope of this paper, only examiners' open-ended feedback is reported here (for the full questionnaire results, see Nakatsuhara *et al*., 2016).

Verbal reports by examiners on the rating of test-takers' performances. After completing all tests each day, examiners took part in verbal report sessions. Seven video-conferencing and seven face-to-face video-recorded tests by the same seven test-takers were selected for this purpose. To cover a range of performance levels, the overall band scores of

[5] The public version of the IELTS Speaking band descriptors is available at https://www.ielts.org/about-the-test/how-ielts-is-scored

the seven test-takers ranged between IELTS Bands 4.5 - 7.5 in one or both of the delivery modes.

The same IELTS examiners participated in the verbal report sessions, and one of the researchers who observed the live tests acted as a facilitator. A single verbal report per test session was collected from the examiner, who interviewed the test-taker. In total, 14 verbal reports were collected. The examiners were first given a short tutorial introducing the procedures for verbal report protocols. Then, following May (2011), verbal report data were collected in two phases, using stimulated recall methodology (Gass & Mackey, 2000):

- Phase 1: Examiners watched a video without pausing while looking at the written comments they made during the live sessions, and made general oral comments about a test-taker's overall task performance.
- Phase 2: Examiners watched the same video clip again, and paused the video whenever they wished to make comments about any features they found interesting or salient related to the four analytic rating categories and any similarities and differences between the two test delivery modes. The researchers also paused the video and questioned the examiners, whenever necessary.

The order of verbal reporting sessions on video-conferencing and face-to-face videos for the four examiners was counter-balanced. The researchers took notes during the verbal report sessions, and all sessions were audio-recorded.

Data analysis

Score analysis. Scores awarded under face-to-face and video-conferencing conditions were compared using both Classical Test Theory (CTT) analysis with the paired samples t-tests (SPSS ver.22), and many-facet Rasch analysis (MFRM) using FACETS 3.71.2 (Linacre, 2013a). From the outset, the CTT analysis was seen as the primary quantitative analysis procedure to address RQ1 because of the constraints imposed by the data collection (i.e., single-ratings). The paired samples t-tests[6] (CTT) were used to examine whether there were any statistically significant differences between the two test-delivery conditions (RQ1). The MFRM analysis was carried out to add further insights into the impact of delivery mode on the scores, and to investigate any inconsistency in different facets of the test. It has to be highlighted here that the results of both the CTT and MFRM analyses are only indicative, given the exploratory nature of this study with a small sample size.

Two MFRM analyses were carried out: a 5-facet analysis with *test-taker ability, rater severity, test version, test mode*, and *rating scale* as facets, and a 4-facet analysis with *test-taker ability, rater severity, test version*, and *rating scale* as facets. This allowed for investigation of the effect of delivery mode on scores in the 5-facet analysis, and for investigation of the performance of each analytic rating scale in each mode as a separate 'component' in the 4-facet analysis. In the 5-facet analysis, only four rating scales were designated as components (i.e., *fluency, lexis, grammar*, and *pronunciation*). In the 4-facet analysis, delivery mode on its own was not designated as a facet, since each of the analytic rating scales in each mode was treated as a separate component, resulting in eight component scores (*face-to-face fluency, video-conferencing fluency*, etc.).

It is necessary here to specify how sufficient connectivity was achieved. As noted above, there was no overlap in the design between examiners and test-takers, resulting in

[6] Although the data distributions indicated slight non-normality, parametric tests were thought to be more appropriate to avoid potential Type 2 errors, given the purpose of this research (N. Verhelst, personal communication, May 6, 2016).

disjoint subsets and insufficient connectivity for a standard MFRM analysis. One way to overcome disjoint subsets is to use group anchoring to constrain the data to be interpretable within a common frame of reference (Linacre, 2013b). Group anchoring involves anchoring the mean of the groups appearing as disjoint subsets; in this case test-takers were grouped according to the examiner by whom they were rated. Group anchoring allows sufficient connectivity for the other facets to be placed onto the common scale within the same measurement framework to be compared on the same Rasch logit scale. Nevertheless, this anchoring method also entails a limitation in that it assumes that the groups are in effect equivalent. Although the assumption of equivalence is largely borne out by the very close mean raw scores for the four groups, the small scale of this study limits the generalisability of the results. The common frame of reference was further constrained by anchoring the difficulty of the test versions. The assumption of test versions being equivalent was borne out by identical observed score means of 6.66 in both Versions 1 and 2, in addition to their equivalence in terms of task design and operationalization (see Nakatsuhara *et al*., 2016 for more details about this anchoring method).

Functional analysis. All 32 recordings were analyzed for language functions elicited from test-takers, using a modified version of O'Sullivan *et al.*'s (2002) checklist, which included 30 language functions under three major functional categories, i.e., informational, interactional, and managing interaction (for the full inventory of the 30 functions and the details of modifications, see Nakatsuhara *et al*., 2016). Although the checklist was originally developed to analyze language functions elicited in paired speaking tasks of the Cambridge General English examinations, the potential to apply it to other speaking tests, including the IELTS Speaking test, has been demonstrated (Brooks, 2003; Inoue, 2013). Three researchers who are familiar with the checklist coded the elicited language functions. Although the three researchers had extensive experience in using O'Sullivan *et al.'s* checklist, they were re-standardized for the use of the checklist modified for this study. The three researchers firstly coded four performances (i.e. 6.25% of the entire data) together. Any discrepancies arisen in their coding results were discussed until agreement was reached for every single case. A few more minor modifications to the checklist were also made at that stage. The remaining dataset was then divided into three groups and coded by one of the three researchers independently. However, for any uncertainties that occurred while coding, a consensus was reached among them.

Following the approach of O'Sullivan *et al.* (2002) and Brooks (2003), the coding was conducted to determine whether each function was elicited in each part of the test, rather than how many instances of each function were observed. The researchers also noted salient and typical ways in which each language function was elicited under the two test conditions in order to enable transcription and detailed analysis of relevant parts of the speech samples. The numbers of test-takers who used each language function in the face-to-face and video-conferencing-delivered tests were then compared using the McNemar's chi-square test (RQ2).

Thematic analysis of examiner comments/verbal reports and researchers' field notes. Unlike the deductive approach to analyzing language functions described above, the remaining data sources were analyzed using an inductive approach (Cohen *et al.*, 2000). The small sample size of this study allowed for the thematic analysis to be carried out in Microsoft Excel.

When the researchers observed live test sessions, they noted similarities and differences in examiners' behavior as interlocutors. These observation notes were typed out and organised in spreadsheet format according to test part and test mode. The notes were then thematically analyzed in conjunction with the examiners' questionnaire responses related to

their behavior as interlocutors and to the examiners' verbal reports. Detailed coding schemes were developed while analyzing the typed data, and the notes in spreadsheet format were coded and classified according to different main themes and sub-themes (RQ3).

All written comments provided by the examiners were compared across the face-to-face and video-conferencing conditions. The researchers who facilitated the 14 verbal report sessions took detailed observational notes and recorded examiners' comments. Resource limitations made it impossible to transcribe all 14 verbal report sessions. Instead, the audio recordings were reviewed to identify key topics and perceptions referred to by the examiners during the verbal report sessions. These topics and comments were then captured in spreadsheet format so they could be coded and categorized according to the different themes that emerged. The thematic content of written commentaries and verbal reports were then qualitatively compared between the face-to-face and video-conferencing modes. The open-ended examiner feedback regarding examiners' perceptions towards the two different delivery modes was analyzed, in conjunction with the results from the analysis of the observation notes, examiners' written comments and verbal reports as described above (RQ3 and RQ4).

These thematic analyses were initially carried out by two of the project researchers who discussed every derived category. Once all analyses were completed, all thematic categories and coded information in Excel tables were presented in a full-day meeting with all the project researchers, and the emerged themes and coding accuracy across different data sources were confirmed.

The results obtained in the analyses of test-takers' scores, linguistic output, examiners' questionnaire responses, written comments, verbal reports, and researchers' field notes were triangulated to explore and give detailed insights into how the video-conferencing-delivery mode compares with the face-to-face mode.

RESULTS

As the 64 tests carried out with the 32 test-takers were perfectly counter-balanced in terms of the order of the two delivery modes and the order in which the two test versions were used by the four examiners, it can be assumed that any order effects or examiner effects were minimized and have not affected the results.

Score analysis

Both CTT analysis and MFRM analysis were carried out to answer *RQ1 (Are there any statistically significant differences in test-takers' scores?).*

Classical Test Theory Analysis. Table 1 shows that there were no significant differences in test scores awarded to the four analytic rating categories and two overall scores (i.e., mean and rounded). We can also note that descriptive statistics indicated slightly lower scores under the video-conferencing condition with the exception of *grammar*, although the mean differences were negligibly small.

INSERT TABLE 1 HERE

Many-faceted Rasch Analysis. The measurement reports for examiners, delivery modes, and rating scales in both the 5- and 4-facet analyses were examined. These reports showed the

severity and difficulty of items (scoring components) within each facet, and the Infit Mean Square index, which is commonly used as a measure of fit in terms of meeting the assumptions of the Rasch model (see Tables 2 and 3).

INSERT TABLE 2 HERE
INSERT TABLE 3 HERE

Infit mean square values were inspected for detecting any underlying inconsistency in each element. Infit values in the range of 0.5 to 1.5 are productive for measurement (Wright & Linacre, 1994), and the commonly acceptable range of Infit is from 0.7 to 1.3 (Bond & Fox, 2007). Infit values for all the examiners, the delivery modes and the rating scales fell within the acceptable range, except *face-to-face lexis*, which was slightly overfitting (i.e., Infit Mnsq=1.33; see Table 3), indicating that the scores given to *face-to-face lexis* were too predictable. Overfit is not productive for measurement but it does not distort or degrade the measurement system. The lack of misfit in the examiner facet indicates that the four examiners did not exhibit inconsistent rating patterns. Furthermore, the lack of misfit across eight rating scales in the 4-facet analysis (see Table 3) is associated with unidimensionality[7] (Bonk & Ockey, 2003) and by extension can be interpreted as indirect evidence that both delivery modes are in fact measuring the same construct.

Of most importance for answering RQ1 are the results for the delivery mode facet in the 5-facet analysis, as shown in Table 2. While video-conferencing was marginally more difficult than the face-to-face mode, the fixed chi-square statistic, which tests the null hypothesis that all elements of the facets are equal, indicated that the two modes were not statistically different in terms of difficulty (X^2=1.8, p=0.19). This reinforces the results of the CTT analysis and strengthens the suggestion that there is no significant impact of delivery mode on scores.

The 4-facet analysis further supports the results of the CTT analysis in that *video-conferencing fluency* and *video-conferencing pronunciation* were the most difficult scales. Although the rating scale facet did not show statistically significant differences (X^2=10.5, p=0.16; see Nakatsuhara *et al*. (2016) for more details), the scales for *fluency* and *pronunciation* seemed to indicate some possible interaction with delivery mode. As will be discussed later, the fact that *pronunciation* was slightly more difficult in the video-conferencing mode seems to relate to the issues with sound quality noted by examiners. For *fluency*, there seemed to be a tendency (at least in some examiners) to constrain verbal back-channeling in the video-conferencing mode, which might have resulted in slightly lower *fluency* scores under the video-conferencing condition.

Language function analysis

Having established score comparability between the two delivery modes, we now turn to the analysis of language functions elicited in the two conditions, to answer *RQ2 (Are there any differences in linguistic output, specifically types of language function, elicited from test-takers?)*

Out of the 30 language functions examined, the majority of functions showed a similar distribution in the two modes, including advanced language functions (e.g., *speculating,*

[7] Additionally, dimensionality in the two delivery modes was examined by correlating scores given to the two modes of the test. Correlation coefficients obtained by Spearman's rho tests (N=32) were 0.829 for *fluency*, 0.827 for *lexis*, 0.739 for *grammar*, 0.848 for *pronunciation*, and 0.914 for the overall scores, which were all significant at a 0.001 level.

elaborating, justifying opinions; for full descriptive and inferential statistics, see Nakatsuhara *et al*., 2016). The observed functions also covered all the 14 functions listed in the IELTS Score Guide (UCLES, 2015:71) as regularly occurring functions in a test-taker's output language[8], although some functions were used by only a few test-takers. This is in line with the IELTS Speaking test design and provides evidence for the comparability of the two modes. There were, however, three language functions that test-takers used significantly differently between the two test modes, as shown in Table 4. These differences emerged only in Parts 1 (Introduction and interview task) and 3 (Examiner and test-taker discussion task) of the speaking test. There were no significant differences in Part 2 (Test-taker long turn task), indicating that the two delivery modes generated comparable functional language in this task type. This is altogether unsurprising, considering the limited co-construction between examiners and test-takers in this task.

INSERT TABLE 4 HERE

More test-takers *asked for clarification* in Parts 1 and 3 of the test under the video-conferencing condition. This is congruent with the examiners' questionnaire feedback in which they indicated that they did not always find it easy to understand each other due to the sound quality of the video-conferencing tests.

The functions *comparing* and *suggesting* were used more often under the face-to-face condition. As expressed in test-takers' interviews (see Nakatsuhara *et al*., 2016), some of them thought that relating to the examiner in the video-conferencing test was not as easy as it was in the face-to-face test, possibly due to the time lag. The immediate nature of the face-to-face condition might have facilitated the test-takers' use of these functions more. A representative sample of transcripts illustrating the three language functions is presented in Nakatsuhara *et al*. (2016).

Analysis of field notes, verbal reports and open-ended comments

We now present results on examiners' test administration behavior (RQ3) and rating behavior (RQ4) under the two delivery conditions. The discussion will be based on an analysis of four different sources of data: retrospective verbal report sessions with examiners; observers' field notes; examiners' written comments; and examiners' feedback questionnaires.

RQ3: Are there any perceived differences in examiners' test administration behavior?

Two main thematic categories emerged:

i) differences reported by examiners in their interaction with test-takers under the face-to-face and video-conferencing condition, and
ii) issues that are perceived as specific to administering a speaking test via a video-conferencing-delivered mode.

[8] The 14 functions are: *providing personal information, expressing a preference, providing non-personal information, comparing, expressing opinions, summarizing, explaining, conversation repair, suggesting, contrasting, justifying opinions, narrating and paraphrasing, speculating,* and *analyzing.*

Differences in examiners' behavior as interlocutors between the two modes. Examiners commented on differences in the following aspects of their own interactional behavior:

- role and frequency of examiner responses
- rate and articulation of examiner speech
- effect of examiner intonation
- use of gestures by examiners (and awareness of gestures used by test-takers)
- issues related to turn-taking management
- requests for clarification

Some examiners were observed to use *verbal and non-verbal response tokens* differently between the two conditions. In the verbal report sessions, one examiner reported that he used more nodding and back-channeling to signal his understanding of what test-takers said in the face-to-face mode to facilitate the test-taker speech, whereas two examiners reported that they used more nodding in the video-conferencing mode for exactly the same reason. It seems that nodding and back-channeling could be deliberately constrained by examiners to avoid video transmission delay; alternatively, they could be deliberately used by examiners as an interactional strategy to compensate for the lack of physical proximity in the video-conferencing condition.

The video-conferencing condition also appeared to lead to a slower speech rate and clearer articulation by examiners, to ensure that test-takers understood them, and possibly to mitigate any perceived technical challenges (e.g., transmission delay or poor sound quality). One examiner commented that as a consequence of her slower speech under the video-conferencing condition, she might have used fewer questions in Part 3 in that mode. However, despite this subjective impression, there was no statistically significant difference in the total number of questions used by examiners between the two modes ($Z(31)=-1.827$, $p=0.068$).

One examiner also referred to *intonation*, suggesting an awareness of the potential for subtleties of tone and implicature to be distorted when the interaction takes place via video-conferencing rather than face-to-face.

A number of examiners' comments concerned *gestures and body language*, indicating that they were sensitive to the potentially differential impact of gesture, movement and body language in the interaction across the two conditions. The consistent message was that in the video-conferencing condition examiners found it harder to use natural gestures as part of their own interaction, and to perceive/read similar gestures when these are used by the test-taker. Limited eye contact in the video-conferencing condition may have also resulted in limited rapport with the test-taker. Furthermore, some simple gestures which routinely support the smooth administration of the face-to-face test, such as finger-pointing the long-turn instruction as he/she explains and showing his/her palm when saying "please start talking", may simply not be possible in the video-conferencing condition.

There were also a number of comments on the challenges posed by the video-conferencing condition for the *management of turn-taking,* highlighting the increased challenge for turn management posed by the video-conferencing condition. It seemed more difficult to signal the examiner's turn initiation, and to determine when the test-taker's turn was complete in the video-conferencing mode. This potentially slows down the turn-taking rate and may result in a reduced language sample being gathered within the time available.

One examiner reported receiving more *clarification questions* in the video-conferencing condition, an observation that was confirmed by the results of the functional analysis discussed earlier.

Issues specific to administering a video-conferencing-delivered speaking test. In their verbal reports, examiners raised a number of issues which they felt impacted negatively on their own role as the facilitator in the test as well as on the smooth running of the speaking test. Three aspects were of particular interest:

- the negative effects of delayed video transmission
- the way the test-taker can impact on the sound quality
- the need to control the direction of the interview.

Related to the above-mentioned turn-management issue, examiners noted how *delayed transmission* in the video-conferencing condition made it difficult to stop a test-taker from continuing to talk, or to intervene and help a test-taker if needed. Two examiners commented that in the video-conferencing condition *the test-taker sometimes impacted negatively, albeit unintentionally, on the sound quality* of their own speech through uncontrolled gestures obscuring the mouth. One examiner noted the stronger need to *control the direction of the interview* in the video-conferencing mode through provision of more language support to prevent the test-taker from going off topic (e.g., adding some examples to her questions; inserting explicit discourse markers such as "let's move on" before introducing a new topic).

Analysis of observers' field notes provided corroborating evidence for all the issues and interactional features discussed above that relate to the examiners' test administration behavior across the two conditions.

RQ4: Are there any perceived differences in examiners' rating behavior?

The verbal report data contained insightful comments from examiners regarding their experience of rating test-takers' spoken performance in the two conditions. The comments suggested that *sound quality* might have impacted on examiners' confidence in rating speech output in a consistent manner. Occasional poor sound quality in the video-conferencing condition seems to have forced examiners to allocate extra resources to certain aspects of the interaction (e.g., careful listening, coping with delayed transmission), possibly at the expense of their attentional capacity for the actual activity of rating. A theme in their comments was the perceived negative impact of poor sound quality in the video-conferencing condition on examiners' judgements of *pronunciation* and *grammar*; the impact seemed less pronounced where judgements of *lexis* are concerned, and there were no explicit comments regarding the *fluency* criterion.

Additional data linked to rating activity were available from examiners in the notes they entered onto the mark sheets during rating. The notes showed that examiners had no difficulty recording evidence for each of the four criteria across the two conditions. However, the notes in the video-conferencing condition make regular reference to technical difficulties and the problem of hearing clearly enough to make the necessary judgement, typically where *pronunciation* was concerned.

The extended responses from examiners provided explanation for some of their comments and were also consistent with themes concerning test administration raised by them in their verbal reports (see Nakatsuhara *et al*., 2016). Analysis of the verbal report data also highlighted some perceptions which might have useful implications for the future use of video-conferencing-delivered speaking tests. For example, exposure to computer-based testing of speaking (through appropriate training and more experience in daily life) develops familiarity and confidence with that mode, which could impact on test-taker and examiner behavior.

DISCUSSION AND FUTURE RESEARCH

This study has carried out a preliminary exploration and comparison of test-taker and examiner behavior across two different delivery modes for the same L2 speaking test, i.e. the standard face-to-face and video-conferencing modes. The quantitative and qualitative data and the different methods of analysis have provided diverse and supplementary pictures of the two delivery modes of the test, which allows for a more in-depth and comprehensive set of findings. Table 5 summarizes the findings for each of the four research questions of this research.

INSERT TABLE 5 HERE

To summarize, the results of this exploratory study comparing face-to-face and video-conferencing-delivery modes in a speaking test context suggest that while the two modes generated similar test scores and a generally comparable range of language functions, some differences were observed in test-takers' functional output and examiners' behavior as both raters and interlocutors. Overall, the findings confirm the score equivalence between the two modes reported by Clark and Hooshmand (1992), Craig and Kim (2010) and Kim and Craig (2012). Interestingly, in the available literature, and in this study, a consistent trend was observed for the video-conferencing mean scores to be slightly lower, although that difference was not in any of the cases statistically significant. This consistent trend warrants further scrutiny, since it has potential implications from a scoring validity perspective. Such future research into scoring validity parameters is also likely to benefit from looking into examiner perceptions of the characteristics of the two modes. The findings of this study indicated that technical aspects such as delayed video in the video-conferencing mode and sound quality led to somewhat different experiences for examiners in the two modes.

In terms of the context validity focus of this study, we extended the research into an investigation of speech functions, which has not been the focus of previous research, and which indicates some differences in functional language across the two modes. More specifically, the functions of comparing and suggesting were used by more test-takers in the face-to-face mode, and asking for clarification were used by more test-takers in the video-conferencing mode.

Overall, the findings present validity evidence supporting the use of the video-conferencing mode as a parallel alternative to the standard face-to-face mode. They also alert us to the complex nature of contextual and scoring validity characteristics and the need to further explore task features such as functional language, interlocutor strategies and rating considerations.

Some of the differences between the two modes may have been affected by the degree of examiner familiarity with video-conferencing delivery. As such, it is recommended that before any decisions are made about deploying an online video-conferencing system as an alternative mode of delivery to an existing face-to-face test or as a format in its own right, further research is required to focus on a range of issues which have remained beyond the scope of this small-scale investigation.

Future studies could examine discourse features such as examiners' and test-takers' rate of speech and test-takers' length of responses. Such an exploration would contribute further evidence for the context validity of the video-conferencing mode. Future studies could also focus on a range of conversational features, such as length of turn, turn interruptions/overlaps, gaps between turns. These features have played a role in interaction, both in the pragmatics and conversation analysis literature (e.g., Itakura, 2001; Sacks, Schegloff, & Jefferson, 1974) and in the L2 testing literature (e.g., Berry, 2007; Brown, 2003;

Galaczi, 2014; Lazaraton, 2002; Nakatsuhara, 2013). It is important, therefore, to focus further investigations on interactional features in the two modes to ascertain whether their use, and therefore their impact on scores, differs across the two modes.

An investigation of cognitive validity considerations was beyond the scope of this study, but it would be a useful area of future research. As argued in Weir's (2005) socio-cognitive validity framework, a core component of validity is evidence on whether the cognitive processes required to complete the task(s) are appropriate, considering the target language use domain. It would be valuable, therefore, to investigate what differences occur between the video-conferencing and standard face-to-face modes in terms of cognitive processing. Such differences might have implications for the construct underlying the video-conferencing mode.

An important issue was the perception of some test-takers that sound quality affected their performance, when the sound and transmission were, in fact, both adequate (as confirmed by a technical advisor who monitored all the sessions in real time). By systematically comparing test-takers' and examiners' perceptions and researchers' field notes, future research can explore when this is more likely to happen. From a practical perspective, the findings will be useful for future examiner training and quality control procedures. Empirically, such an exploration will contribute insights to test-taker perceptions of the contextual validity characteristics of the video-conferencing mode.

Comments from examiners and test-takers also pointed to the need for explicit examiner and test-taker training if the introduction of video-conferencing oral testing is to become operational in the future, in order to minimize any possible disadvantages of this format. Since the completion of this study, such an endeavor has already been initiated as part of a follow-up project, and materials to prepare test-takers for video-conferencing communication and examiner training materials for video-conferencing-specific techniques have been developed. The follow-up project is a larger-scale replication study which aims to confirm and add generalizability to the findings reported in this paper, after minimizing the effects of video-conferencing familiarity on examiners and test-takers. Since the small sample size of this study limited our ability to explore score differences fully, the follow-up project is also designed to allow for more powerful and accurate statistical analysis of differences in scores awarded on the two different modes of oral tests, leading to more generalizable conclusions.

Finally, the effect of the technical issues in deploying the video-conferencing mode in research or operational settings should not be underestimated. Although Zoom, the computer-mediated communication software employed in this study, is generally considered to be more stable than other similar programs, some technical issues of sound quality and delayed video transmission were encountered. This is an issue that must be carefully considered and addressed in any operationalization of video-conferencing systems in the context of L2 speaking assessment and it is critically important that the test venues have stable internet connection, sufficient bandwidth and local technical support[9]. As such, we recommend that the video-conferencing mode should not be seen as a replacement for the standard face-to-face mode of the test investigated here, but should be treated as a viable alternative which is empirically supported by validation evidence, but which may be difficult to currently deploy in large-scale operational conditions.

As we saw in the review of the literature at the beginning of this paper, to date there is little research into the effect of a video-conferencing mode of a speaking test (Clark &

[9] It should be noted that, in this exploratory study, transmission demands were low since examiner rooms and test-taker rooms were located in the same building. Further technical trials with a bespoke platform for the video-conferencing mode are in progress with geographically distant examiners and test-takers.

Hooshmand, 1992; Craig & Kim, 2010; Kim & Craig 2012), and no research into the use of this mode on a par with the standard face-to-face mode in a high-stakes test context. The research reported here adds to a small body of empirical work focused on whether technology-mediated delivery of a face-to-face speaking test is testing the same or a different speaking construct as its face-to-face counterpart. The results are not, at this stage, conclusive, but nevertheless present a multi-faceted view of this new test mode which focuses on a range of important validity concerns. It is hoped that as technology keeps improving and becomes more accessible, a growing body of literature will become available to provide more empirical information about this exciting new format in L2 speaking assessment.

REFERENCES

Abrams, Z. I. (2003). The effect of synchronous and asynchronous CMC on oral performance in German. *The Modern Language Journal, 87*(2), 157-167.

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bernstein, J., Van Moere, A., & Cheng, J. (2010). Validating automated speaking tests. *Language Testing, 27*(3), 355-377.

Berry, V. (2007). *Personality differences and oral test performance*. Frankfurt am Main: Peter Lang.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model. Fundamental measurement in the human sciences (2nd ed.)*. Mahwah, NJ: Lawrence Erlbaum.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20*(1), 89-110.

Brooks, L. (2003). Converting an observation checklist for use with the IELTS Speaking Test. *Cambridge ESOL Research Notes, 11,* 20-21.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*(1), 1-25.

Clark, J. L. D. (1988). Validation of a tape-mediated ACTFL/ILR-scale based test of Chinese speaking proficiency. *Language Testing, 5*(2), 197-205.

Clark, J. L. D., & Hooshmand, D. (1992). 'Screen-to-screen' testing: An exploratory study of oral proficiency interviewing using video conferencing. *System, 20*(3), 293-304.

Cohen, J. (1988) *Statistical power analysis for the behavioural sciences (2nd edition)*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Craig, D. A., & Kim, J. (2010). Anxiety and performance in videoconferenced and face-to-face oral interviews. *Multimedia-assisted Language Learning, 13*(3), 9-32.

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research (2nd edition).* Thousand, Oaks, CA: Sage Publications.

Field, J. (2011). Cognitive validity. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking* (Studies in Language Testing, Vol. 30). (pp. 65-111). Cambridge: Cambridge University Press.

Galaczi, E. D. (2010). Face-to-face and computer-based assessment of speaking: Challenges and opportunities. In L. Araújo (Ed.), *Computer-based assessment of foreign language speaking skills* (pp. 29-51). Luxemburg: European Union.

Galaczi, E. D. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics, 35*(5), 553-574.

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Lawrence Erlbaum.

Hoejke, B., & Linnell, K. (1994). Authenticity in language testing: Evaluating spoken language tests for international teaching assistants. *TESOL Quarterly, 28*(1), 103-126.

Inoue, C. (2013 November). *Investigating the use of language functions for validating speaking test specifications.* Paper presented at Language Testing Forum 2013, Nottingham, UK.

Isaacs, T. (2010). *Issues and arguments in the measurement of second language pronunciation.* Unpublished PhD thesis, McGill University, Montreal.

Itakura, H. (2001). Describing conversational dominance. *Journal of Pragmatics, 33*(12), 1859-1880.

Kenyon, D., & Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other proficiency assessments. *Language Learning and Technology, 5*(2), 60-83.

Kiddle, T., & Kormos, J. (2011). The effect of mode of response on a semidirect test of oral proficiency. *Language Assessment Quarterly, 8*(4), 342-360.

Kim, J., & Craig, D. A. (2012). Validation of a videoconferenced speaking test. *Computer Assisted Language Learning, 25*(3), 257-275.

Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests* (Studies in Language Testing, Vol. 14). Cambridge: UCLES/Cambridge University Press.

Linacre, M. (2013a). *Facets computer program for many-facet Rasch measurement, version 3.71.2.* Beaverton, Oregon: Winsteps.com.

Linacre, M. (2013b). *A user's guide to FACETS: Rasch-model computer programs*, available on line at www.winsteps.com/a/facets-manual.pdf..

Luoma, S. (1997). *Comparability of a tape-mediated and a face-to-face test of speaking: A triangulation study.* Unpublished Licentiate thesis, University of Jyvaskyla, Jyvaskyla. Retrieved May 14, 2014 from http://urn.fi/URN:NBN:fi:jyu-1997698892.

May, L. (2011). *Interaction in a paired speaking test: The rater's perspective.* Frankfurt am Main: Peter Lang.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA & Oxford: Blackwell.

Nakatsuhara, F. (2013). *The Co-construction of conversation in group oral tests*. Frankfurt am Main: Peter Lang.

Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. (2016). Exploring performance across two delivery modes for the same L2 speaking test: Face-to-face and video-conferencing delivery - A preliminary comparison of test-taker and examiner behaviour, *IELTS Partnership Research Papers, 1,* 1-67.

Nakatsuhara, F., Inoue, C., & Taylor, L. (2017). An investigation into double-marking methods: comparing live, audio and video rating of performance on the IELTS Speaking Test. *IELTS Research Report Online Series, 1,* 1-49.

O'Loughlin, K. (2001). *The equivalence of direct and semi-direct speaking tests* (Studies in Language Testing, Vol 13). Cambridge: Cambridge University Press.

O'Sullivan, B., Weir, C. J., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing, 19*(1), 33-56.

O'Sullivan, B., & Weir, C. J. (2011). Test development and validation, In B. O'Sullivan (Ed.), *Language testing theories and practices* (pp. 13-32). Basingstoke: Palgrave Macmillian.

Qian, D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers *Language Assessment Quarterly, 6*(2), 113-125.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*(4), 696-735.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing, 11*(2), 99-123.

Smith, B. (2003). Computer-mediated negotiated interaction: An expanded model. *The Modern Language Journal, 87*(1), 25-38.

Stansfield, C. (1990). An evaluation of simulated oral proficiency interviews as measures of oral proficiency. In J. E. Alatis (Ed.), *Georgetown University Roundtable of Languages and Linguistics 1990* (pp. 228-234). Washington, D.C.: Georgetown University Press.

Stansfield, C., & Kenyon, D. (1992). Research on the comparability of the Oral Proficiency Interview and the Simulated Oral Proficiency Interview. *System, 20*(3), 347-364.

Taylor, L. (2011). Introduction. In L. Taylor (Ed.) *Examining speaking: Research and practice in assessing second language speaking* (Studies in Language Testing, Vol. 30). (pp. 1-35). Cambridge: Cambridge University Press.

Taylor, L., & Galaczi, E. D. (2011). Scoring validity. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking* (Studies in Language Testing, Vol. 30). (pp. 171-233). Cambridge: Cambridge University Press.

UCLES (2015). *IELTS Scores Guide*. Cambridge: UCLES.

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing, 29*(3), 325-344.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.

Weir, C. J., Vidakovic, I., & Galaczi, E. (2013). *Measured constructs* (Studies in Language Testing, Vol. 37). Cambridge: Cambridge University Press.

Wright, B., & Linacre, M. (1994). *Reasonable mean-square fit values*. Retrieved March 27, 2012 from www.rasch.org.

Yanguas, I. (2010). Oral computer-mediated interaction between L2 learners: It's about time! *Language Learning and Technology, 14*(3), 72-93.

Table 1: Differences in test scores by paired samples t-test

Rating Category	Test Mode	Mean	SD	Mean Difference	t (d.f.=31)	Sig. (2-tailed)
Fluency	face-to-face	6.594	1.043	.188	1.791	.083
	video-conferencing	6.406	1.160			
Lexis	face-to-face	6.750	1.047	.0313	.297	.768
	video-conferencing	6.719	1.143			
Grammar	face-to-face	6.625	1.008	.000	.000	1.000
	video-conferencing	6.625	1.100			
Pronunciation	face-to-face	6.688	0.780	.156	1.717	.096
	video-conferencing	6.531	0.879			
Overall[10] (mean)	face-to-face	6.664	0.829	.094	1.459	.155
	video-conferencing	6.570	0.982			
Overall (rounded)	face-to-face	6.547	0.883	.078	1.044	.305
	video-conferencing	6.469	0.991			

Table 2: Delivery mode measurement report (5-facet analysis)

Test Mode	Logit Measure	Standard Error	Infit Mean Square
face-to-face	-0.16	0.17	1.08
video-conferencing	0.16	0.17	0.85
(Population): Separation 0.00 Strata 0.33 Reliability (of separation) 0.00			
(Sample): Separation 0.87 Strata 1.49 Reliability (of separation) 0.43			
Model, Fixed (all same) chi-square: 1.8 d.f.: 1 significance (probability): 0.19			

Table 3: Rating scale measurement report (4-facet analysis)

Rating Category	Test Mode	Logit Measure	Standard Error	Infit Mean Square
Fluency	face-to-face	0.13	0.34	1.01
	video-conferencing	0.71	0.34	0.82
Lexis	face-to-face	-0.57	0.35	1.33
	video-conferencing	-0.45	0.35	0.79
Grammar	face-to-face	0.02	0.34	0.93
	video-conferencing	0.02	0.34	0.96
Pronunciation	face-to-face	-0.22	0.34	1.06
	video-conferencing	0.36	0.34	0.83
(Population): Separation 0.57 Strata 1.09 Reliability (of separation) 0.24				
(Sample): Separation 0.71 Strata 1.28 Reliability (of separation) 0.34				
Fixed (all same) chi-square: 10.5 d.f.: 7 significance (probability): 0.16				

[10] The first overall category shows mean overall scores, and the second overall category shows overall scores that are rounded down as in the operational IELTS test (i.e., 6.75 becomes 6.5, 6.25 becomes 6.0).

Table 4: Differences in elicited language functions by McNemar's chi-square test[11]

[Part] Function	Test Mode	Count	Mean	SD	X^2 (d.f.=1)	Sig. (2-tailed)
[Part 1] asking for clarification	face-to-face	4	.13	.34	1.723	.000
	video-conferencing	19	.59	.50		
[Part 3] comparing	face-to-face	29	.91	.30	12.565	.012
	video-conferencing	20	.63	.49		
[Part 3] suggesting	face-to-face	12	.38	.49	12.565	.012
	video-conferencing	3	.09	.30		
[Part 3] asking for clarification	face-to-face	15	.44	.50	1.195	.022
	video-conferencing	24	.72	.46		

Table 5: Summary of findings

Research Questions Comparing the standard face-to-face and video-conferencing mode:	**Findings**
RQ1: Are there any statistically significant differences in test-takers' scores?	Although scores for *fluency, lexis* and *pronunciation* were slightly lower in the video-conferencing mode, actual score differences were negligibly small and none of these differences was statistically significant.
RQ2: Are there any differences in linguistic output, specifically types of language function, elicited from test-takers?	No significant differences were observed in many of the language functions. Significant differences were found in *comparing* and *suggesting* in Part 3, which were used by more test-takers in the face-to-face mode. *Asking for clarification* in Parts 1 and 3 was used by more test-takers in the video-conferencing mode.
RQ3: Are there any perceived differences in examiners' test administration behavior?	Examiners reported differences in the use of response tokens, articulation and speed of their speech, intonation, gestures, turn-taking and requests for clarification. Delayed video transmission and sound quality affected their behavior as interlocutor (e.g., difficulty in intervening).
RQ4: Are there any perceived differences in examiners' rating behavior?	Examiners reported that sound quality and delayed video affected the ease of rating as they had to allocate attention to these video-conferencing-specific aspects which are not present under the face-to-face condition. They sometimes found it difficult to rate *pronunciation* and *grammar* under the video-conferencing condition.

[11] Given the purpose of this research, the alpha levels were not adjusted using Bonferroni corrections in order to avoid potential Type 2 errors.

Figure 1: Research design